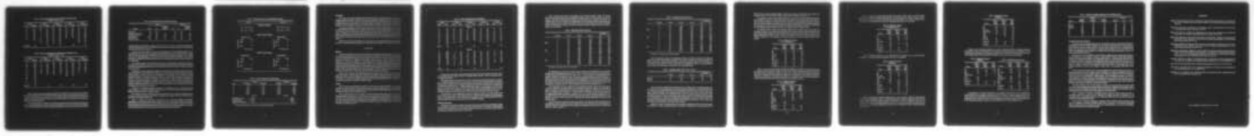
AD-A050 829 AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TEX F/G 5/9
SIMULATED AND EMPIRICAL STUDIES OF FLEXILEVEL TESTING IN AIR FO--ETC(U)
SEP 77 D A HARRIS, R J PENNELL
UNCLASSIFIED AFHRL-TR-77-51 NL

| OF |
AD
A050829



END
DATE
FILMED
4-78
DDC

AFHRL-TR-77-51

AIR FORCE HUMAN RESOURCES LABORATORY

SIMULATED AND EMPIRICAL STUDIES OF FLEXILEVEL TESTING IN AIR FORCE TECHNICAL TRAINING COURSES

By

Dickie A. Harris, Capt, USAF
Roger J. Pennell

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado 80230

September 1977

Final Report for Period 1 May 1975 – 30 April 1977

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE,TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Technical Training Division, Air Force Human Resources Laboratory, Lowry Air Force Base, Colorado 80230, under project 1121, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved for publication.

MARTY R. ROCKWAY, Technical Director
Technical Training Division

DAN D. FULGHAM, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

14

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFHRL-TR-77-51✓

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

6

4. TITLE (and Subtitle)
SIMULATED AND EMPIRICAL STUDIES OF FLEXILEVEL TESTING IN AIR FORCE TECHNICAL TRAINING COURSES.

9

5. TYPE OF REPORT & PERIOD COVERED
Final rept.
1 May [redacted]75 – 30 Apr[redacted]77,

6. PERFORMING ORG. REPORT NUMBER

10

7. AUTHOR(s)
Dickie A. Harris
Roger J. Pennell

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Technical Training Division ✓
Air Force Human Resources Laboratory
Lowry Air Force Base, Colorado 80230

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
16 62205F 17
11210309

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)✓
Brooks Air Force Base, Texas 78235

11

12. REPORT DATE
Sep[redacted]77

13. NUMBER OF PAGES
~~22~~ 15 24p.

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

adaptive testing
computerized testing
flexilevel strategy
simulation studies
technical training

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study used a series of simulations to answer questions raised by empirical studies. The first study showed that for reasonable high entry points, parameters estimated from paper-and-pencil test protocols cross-validated remarkably well to groups actually tested at a computer terminal. This suggested that feasibility studies; i.e., running actual subjects, may not be called for. The second study showed that the proportion correct during flexilevel testing was a sensitive measure of student performance. It was also concluded that the modest time savings (12 to 15 percent) was due to the parameters used to implement flexilevel testing. Study III showed that a 50 percent savings in items, and, potentially, a large savings in test time could be realized through the implementation of alternate flexilevel strategies. In summary, the overall conclusion from the three studies was that flexilevel testing, with → over

404 415

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Item 20 Continued:

variable entry, offers an easily implemented testing procedure with potential for significant dollar savings at minimal risk.

TABLE OF CONTENTS

LIST OF TABLES

SIMULATED AND EMPIRICAL STUDIES OF FLEXILEVEL TESTING IN AIR FORCE TECHNICAL TRAINING COURSES

I. INTRODUCTION

In an environment such as is offered by the Advanced Instructional System (AIS), the potential benefits derivable from adaptive testing become a practical reality. The AIS is an advanced development program to develop a computer based educational and training system for the Air Force. The heart of the system is a CDC Cyber-70 Computer which currently manages the training process for four courses at Lowry Technical Training Center, Lowry AFB, Colorado, through two types of terminals. The type A terminal is an interactive plasma display terminal with graphic capabilities, while the type B management terminal has test form reading and scoring capabilities along with a line printer for issuing student prescriptions. The system is designed to manage the individualized instructional process of a large number of students who spend approximately 33 percent of their time in a testing mode. Thus, with a large student flow through AIS courses requiring extensive testing, considerable payoff in terms of reduced training time is potentially available from procedures which reduce testing times without compromising instructional effectiveness.

Adaptive testing has been investigated under a variety of rubrics such as branched testing, response contingent testing, sequential testing, tailored testing. We shall use the general term adaptive testing to characterize any attempt to match test items to examinees based on a response history, with the goal of reducing testing time, or obtaining more valid and/or more reliable ability estimates.

Background

Realizing the potential of adaptive testing in a system such as the AIS, the Air Force Human Resources Laboratory, Brooks AFB, Texas, initiated a multi-phase research study beginning with the identification of a suitable algorithm to drive an adaptive testing program. During Phase I, the flexilevel approach of Lord (1971a, 1966b) was identified as the tentative algorithm (Hansen, Johnson, Fagan, Tam, & Dick, 1974). Flexilevel testing has a number of advantages over other methods of adaptive testing. Namely, it is easily implemented, it does not require a large item pool, and theoretically it requires only (n+1)/2 items (where n is the number of items in the total test pool) to test each examinee. For example, a 25 item test would require only 13 items to test each examinee. The flexilevel test (Lord, 1971a, 1971b) first administers the item of median difficulty (difficulty levels ascertained from pretesting). If an item is answered incorrectly, the next easiest, unanswered item is given. If an item is answered correctly, the next hardest, unanswered item is given. An examinee continues testing until he has answered (n+1)/2 items.

Phase II of the research consisted of experimental studies conducted in the Inventory Management (IM) and Precision Measuring Equipment (PME) courses. The Block II test of the IM course was used for the implementation of Study I (Hansen, Harris, & Ross, 1977a) while the Block II and Block IV tests of PME were used in Study II (Hansen, Harris, & Ross,1977b). The purpose of Study I was to validate the flexilevel, adaptive testing paradigm with the primary goals of reducing test time. Each student was individually entered in the test, given the flexilevel adaptive test and then all remaining items. This design was employed in order to fulfill the operational requirements of the training system. The results revealed an extremely high part-whole correlation (r = .94) between the flexilevel and total test scores. The flexilevel test, however, required 39.5 percent fewer items with a concomitant time savings of 18.4 percent.

As mentioned, Study II was performed in Blocks II and IV of the PME course. A task analysis was used to group items into five scales and to construct a hierarchy of scales within the test. The intention was to explore the feasibility of adaptively testing both within and across scales. Test validity analyses yielded high part-whole correlations between adaptive test and total test scores (r's = .95). In addition, the time savings associated with adaptive testing approximated 30 percent for both blocks (Hansen, Harris, & Ross, 1977b). Following completion of the two empirical studies several questions concerning the efficacy of adaptive testing remained to be answered.

The purpose of this report is to present the results of three simulation studies designed to answer questions raised by the empirical studies. The first simulation study was designed to evaluate the need for conducting empirical studies. The second simulation study was designed to reconstruct the testing situation and analyze the data for different purposes. And finally, the third study was designed to simulate, using Study II test protocols, the effects of adaptive movement across scales as well as within scales.

II. STUDY I

Objective

The thrust of Study I was to explore the kinds of conclusions which might be made by simulating flexilevel testing on paper-and-pencil protocols and comparing the results (i.e., estimated parameters) to those data actually collected on the computer terminal (Phase II). The intent was to evaluate the extent to which the actual implementation and testing of the model on a computer terminal can be avoided.

A number of simulation studies of adaptive testing have been conducted; among these are Bryson (1972); Cleary, Linn, and Rock (1968a, 1968b); Linn, Rock, and Cleary (1970); and Patterson (1962). These studies have largely been concerned with ascertaining the potential benefits derivable from an adaptive testing paradigm, rather than extrapolating simulated results to actual adaptive data as this study did. Basically, the question posed by the present study was, "Must one actually conduct an empirical study such as that conducted during Phase II to ascertain adaptive testing feasibility?" And furthermore, "To what extent do simulated results parallel results under actual PLATO testing conditions?"

Method

A sample of 186 paper-and-pencil protocols was obtained from Inventory Management/Materiel Facilities (IM/MF) Block II. The test was composed of the same items used in the Phase II experiment. The sample was divided into two equal parts; i.e., a calibration (C) sample and a validation (V) sample. The C sample was used to estimate parameters necessary to implement the flexilevel testing algorithm. These parameters were then validated on the V sample in order to evaluate the stability of various dependent measures. The parameters estimated were the item difficulties, which imply the item ordering for flexilevel presentation, and the regression parameters for converting the flexilevel score into an estimated total score. Admittedly, the flexilevel score could have been used to make the necessary pass/fail decisions required in a criterion-referenced testing situation such as found in Air Force technical training; however, for two reasons it was desirable to translate back to the total score metric (percent correct). First, this is the metric traditionally used to assign scores, and second, the extent to which the flexilevel score reproduces the total score is a prime dependent measure in evaluating the feasibility of flexilevel testing. The flexilevel score was derived as follows: Let A index the set of items taken under flexilevel testing and let d_i, $i \in A$, represent the difficulty of the i-th item expressed as percent of the C sample answering correctly. Further, let

$$s_i = \begin{cases} 1 \text{ if item i answered correctly,} \\ -1 \text{ if item i answered incorrectly.} \end{cases}$$

Then, the flexilevel score for the j-th examinee was defined as

$$F_j = \sum_{i \epsilon A} s_i d_i \qquad (1)$$

Stated more simply, F_j was the sum of the difficulties of items answered correctly minus the sum of the difficulties of items answered incorrectly.

Since the total score, X_j, say, was available as the sum of correct responses divided by the number of items in the item pool, (n = 25), we used the usual regression equation,

$$\hat{X}_j = a + bF_j \tag{2}$$

to estimate the total score and the associated error $|\hat{X}_j - X_j|$. It should be noted that the usual flexilevel rule of administering (n+1)/2 items to each examinee was departed from in both the Phase II study and here. That is, testing for a particular examinee was terminated if he was to take a harder item, but had already answered all of the harder items, or if he was to take an easier item, but had already taken all of these. This decision rule was used because one of the dependent measures was the number of items required to terminate testing as a function of entering examinees at varying locations on the item hierarchy.

The dependent variable analyzed besides those mentioned above (viz, effect of item hierarchy variable entry and error in reproducing total score) was classification error. Here we examined, for a range of criterion levels, the error rate using $\hat{X}_j$ to classify students as failing or passing relative to their known classification based on X_j.

In addition to the C and V samples, a third sample (N = 100) was obtained by randomly selecting test protocols of students who had gone through Phase II testing on the computer. This was possible since at the completion of each flexilevel session (using the same stopping rule described above) all items on the 25 item instrument which had not been administered were given. Thus, complete item protocols were available on this cross-validation (CV) group.

One intention of the Phase II study was to explore the utility of adaptively entering examinees into the item hierarchy. The entry point was calculated using three aptitude tests taken before the students entered training. It was thought that adaptive entry might further reduce testing time over savings attributable to taking only (n+1)/2 items. Unfortunately the CV sample was obtained when monitors were having difficulty obtaining the aptitude scores. Therefore, the majority of the sample was entered at the (n+1)/2-th item.

The comparison of the flexilevel results in the CV group using the parameters estimated in the C group explored whether a feasibility study such as Phase II needs to be conducted. Theoretically, the only difference between the CV and C groups was the use of a computer terminal to administer the test. This assumes item independence in the sense that taking items in a different order would not affect the test score.

Results and Discussion

The item difficulties for the 25 items under study are presented in Table 1. The mean item difficulty, an estimate of the mean test score, was .804. Typically, criterion-referenced test items tend to be quite easy; however, one of the items is exceptionally hard (item 6). Eliminating item 6 raises the mean to about .84, which indicates that about 16 percent of the sample misses an average item. The difficulties in Table 1 implied the ordering of the items for the simulated flexilevel testing, equal item difficulties implied an arbitrary ordering.

Next, the regression parameters for Equation 2 were estimated. Regression estimates for entering the item hierarchy at items 3, 5, 7, 9, 11, 13, and 15 were calculated. These estimates are presented in Table 2 along with the correlation (validity) between X, the total score, and F, the flexilevel score (see Equation 1). The lower down (easier items) on the item hierarchy students were entered, the more items were required to terminate the flexilevel algorithm. This was vividly displayed by the trend of the regression weights. That is, increasing the entry point reduced the constant term, a, and increased the importance of the b term corresponding to the flexilevel score. The validities beginning at entry point 7 were quite good, indicating a high degree of accuracy in predicting total score. However, the cross-validated validities were more interest.

Table 3 presents the V and CV group validities along with the C group for comparison. It should be noted that X_j, the estimated total score was computed using the weights developed in the C group. The

Table 1. **Item Difficulties, Group C**

Item	Difficulty
1	.968
2	.936
3	.819
4	.851
5	.809
6	.468
7	.670
8	.819
9	.819
10	.638
11	.915
12	.777
13	.777
14	.862
15	.894
16	.840
17	.840
18	.840
19	.723
20	.862
21	.691
22	.819
23	.755
24	.926
25	.777

Table 2. **Regression Weights and Validities, Group C**

Entry Point	a	b	Validity
3	.714	.388	.654
5	.656	.509	.773
7	.617	.560	.847
9	.578	.612	.926
11	.555	.631	.952
13	.524	.661	.972
15	.503	.671	.981

Table 3. **Validities by Entry Point**

Entry Point	Group		
	C	V	CV
3	65[a]	75	60
5	77	78	69
7	85	87	79
9	93	93	83
11	95	95	93
13	97	97	96
15	98	98	98

[a] Decimal points omitted.

validities for the V group were strikingly high, in some cases higher than the C group. This indicated that the error in utilizing "nonoptimal" regression weights and item difficulties was essentially non-existent. Some shrinkage was encountered in the CV group. However, this shrinkage all but evaporated after entry point 11. This indicated that parameters developed on paper-and-pencil protocols cross-validate to results obtained by use of computer terminals for high entry levels.

Since the items used to construct the flexilevel score were also used (together with additional items) to compute the total score, the validities reported in Table 3 are inflated to some extent. The total score was computed by summing 1's and 0's corresponding to a correct or incorrect item, whereas the flexilevel score was computed by summing weighted item difficulties. Doubtless, the weighted item difficulties have some built-in minimum correlation with the 1-0 protocol.

Table 4 presents the average percent of items needed to terminate the flexilevel algorithm as a function of entry point. For example, when entering at item 5, all three groups required an average of 30 percent of the total 25 items, or 7.5 items, to terminate the algorithm. The differences between the C sample and the V and CV samples presumably reflect an increase in test items required by using nonoptimal difficulties, and thus a nonoptimal item hierarchy for flexilevel branching. However, this effect was decidedly minimal.

Table 5 presents in terms of number of items, the average, absolute error made in predicting total score. For example, entering at the 11-th item, the estimated total score ($\hat{X}_j$) differs by an average of .9

Table 4. **Percent Items Required to Terminate Testing**

Entry Point	Group V	Group C	Group CV
3	20	20	19
5	30	30	30
7	41	40	41
9	52	50	52
11	62	60	62
13	70	69	72
15	78	77	80

Table 5. **Item Error in Predicting Total Score**

Entry Point	Group V	Group C	Group CV
3	2.0	1.7	1.9
5	1.7	1.5	1.8
7	1.5	1.3	1.4
9	1.2	1.0	1.3
11	.9	.9	.9
13	.7	.6	.7
15	.5	.6	.5

item from the known total score (X_j). Similar to Table 3, these data show comparable results across the three groups entering at item 11 and above.

Table 6 shows the average percentage of error of classification across various criterion levels. For example, for a criterion of .70 if $\hat{X}_j \geqslant .70$ and $X_j \geqslant .70$ or if $\hat{X}_j < .70$ and $X_j < .70$, the j-th student is properly classified. However, if $\hat{X}_j \geqslant .70$ and $X_j < .70$ or if $\hat{X}_j < .70$ and $X_j \geqslant .70$, there has been a classification error relative to the criterion of 70 percent. The percent of these errors averaged over criterion levels .40, .44, ..., .96 is the statistic presented in Table 6. Entering at item 3, the cross-validated percentage of errors is about 11.5 percent which doubtless would be unacceptably high to most course designers. On the other hand, errors of 6 or 7 percent might be acceptable when balanced against the decrease in overall training time.

Table 6. **Classification Error by Entry Point**

Entry Point	Groups V	Groups C	Groups CV
3	14[a]	11	12
5	11	10	11
7	10	8	9
9	8	7	9
11	6	6	7
13	5	5	6
15	4	4	5

[a] Percent misclassified.

Conclusions

Making any decision regarding the implementation of adaptive testing involves a trade off between potential gains vs. potential losses. It has been shown that fairly substantial decreases in required test items are obtainable with very accurate estimation of total score (an empirical question remaining is whether there is a decrease in testing time associated with the decrease in test items). The trade-off is relative to the decision categorizing an examinee incorrectly as passing or failing based on a flexilevel score. The above results indicate that this type of error ranges from about 5 to 12 perecent. It should be noted, however, that the criterion used to gauge this error was the total score; this is a far from ideal criterion. What is needed, of course, is the "true score;" i.e., the unknown indicator of whether a student has accomplished the behavioral objective, imperfectly measured by the total test score, or not. Lacking such an indicator we have used the total score. However, there is no reason why the flexilevel test could not be making the more valid decisions relative to the "true score." Indeed, this is one of the theoretical benefits attributable to adaptive testing.

The foregoing data have indicated that for reasonable high entry points, parameters estimated from paper-and-pencil test protocols cross-validate remarkably well to groups actually tested at a computer terminal using a flexilevel algorithm. This suggests that feasibility studies, running actual subjects, may not be called for. Rather, simulated results based on paper-and-pencil protocols may lead to a quick decision as to whether to implement adaptive testing.

III. STUDY II

Objective

The objectives of Study II were to summarize the data collected under the Phase II contract effort, and to offer some conclusions concerning the efficacy of flexilevel testing in an on-going training environment. The analysis was, of course, constrained by the manner in which the study was implemented; however, the present analysis takes a somewhat different cut at the data.

Method

A sample was obtained of 133 PME students who block tested on a computer terminal. Of those 133 protocols, 61 were Block II tests and 72 were Block IV tests. Both block tests contained 40 items; however, the subject matter covered by the tests was quite different.

A task analysis was done in order to construct a hierarchical structure for each test. The task analysis grouped items into five relatively homogeneous *scales according to item content.* The scales were then placed in a hierarchical structure based on the relationships defined by the task analysis.

All students entered the test at the median difficulty item of the first scale and were presented items based on the flexilevel algorithm described in Study I. After completing the flexilevel portion of each scale, the students were given the remainder of the items and then started at the median difficulty item in the next scale. This procedure was continued until all five scales were completed.

Results

The items comprising the scales along with their difficulties are presented in Table 7 and Table 8. As in Study I, the items were quite easy, the scale mean difficulties ranging from .81 to .94 in Block II and from .81 to .93 in Block IV. Notice, also, that the average difficulty of a scale does not necessarily correspond to the position of the scale within the hierarchy. That is, the scales were ranked in the hierarchy not by average difficulty but rather, by content.

The variables of interest were the proportion of items answered correctly during the flexilevel portion of the test (S_j) and the flexilevel score (F_j), the latter being modified slightly from Study I. Namely, let R be the set of items the student got right, W the set wrong during flexilevel testing, and P_i the difficulty of the i-th item as obtained from Tables 7 and 8. Then:

$$F_j = \sum_{i \epsilon R} (1-P_i) - \sum_{k \epsilon W} P_k \qquad (3)$$

defines the flexilevel score for the j-th student. As well, we shall be interested in the percent of items saved, the amount of time saved relative to taking the full 40 item test, and the remainder score – the score achieved on those items not taken during the flexilevel portion.

Table 7. **Items Comprising Scales and Difficulties for the Block II Test**
(Calibration Sample N = 105)

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
11	.97	24	.97	15	.98	26	.89	34	.95
10	.96	14	.96	29	.94	25	.88	31	.94
6	.96	1	.95	21	.94	39	.88	36	.93
6	.95	5	.90	16	.93	27	.81	37	.90
12	.94	3	.90	20	.92	40	.81	32	.85
7	.92	2	.75	17	.89	28	.70	38	.84
8	.86	23	.74	18	.87			35	.77
		13	.72	19	.85			33	.63
		4	.70	22	.84			30	.51
Mean Diff	.94		.84		.91		.83		.81

Table 8. **Items Comprising Scales and Difficulties for the Block IV Test**
(Calibration Sample N = 113)

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
15	1.00	1	.96	29	1.00	31	.98	38	.96
16	1.00	10	.90	26	.99	39	.88	4	.95
18	1.00	11	.88	24	.98	37	.88	14	.85
8	.96	5	.88	23	.97	34	.87	13	.84
21	.96	22	.82	25	.94	32	.82	28	.81
2	.92	35	.62	27	.83	33	.70	17	.70
19	.86	7	.61	30	.72	36	.69		
12	.81					40	.57		
20	.82								
3	.67								
6	.58								
9	.58								
Mean Diff	.93		.81		.92		.80		.85

Table 9 contains the means, standard deviations and correlations with total score for S_j, F_j, percent items saved, and remainder scores for Blocks II and IV. Both S_j and F_j were almost perfectly related to the total score as evidenced by the correlation of .98. This indicated that after the student had taken about 70 percent of the items in Block II and 75 percent of the items in Block IV, the prediction of his total score from S_j or F_j was almost perfect.

It was surprising that the relatively crude measure, S_j, performed as well as F_j which was intended to be the more sensitive measure. F_j takes into account the difficulty of the item the student takes: passing an item, i, which is relatively easy results in a relatively small increase in score $(1\text{-}P_i)$, and a larger increase for a

Table 9. **Summary Statistics for Dependent Measures**

	Block II			Block IV		
Measure	Mean	SD	Correlations with Total Score	Mean	SD	Correlations with Total Score
Total Score	.85	.39		.82	.39	
S_j (Proportion Correct)	.82	.40	.98	.79	.37	.98
F_j (Flexilevel Score)	.56	.19	.98	.47	.16	.98
% Items Saved	30.4	.89	.96	24.6	.83	.91
Remainder Score	.94	.35	.72	.93	.16	.66

relatively difficult item. Whereas, missing a relatively easy item, i, results in a relatively large decrease in score (P_i), and a lesser decrease for a relatively hard item. However, for the parameters of the present study, both measures performed equally well.

One can notice from Table 9 that the mean remainder score was substantially higher than the corresponding total score. This was expected, since with relatively easy items, students tended to emerge from each scale after taking the most difficult item. Therefore, the remaining items tended to be the easiest items, with an associated higher score. Since the items were relatively uniform in difficulty, S_j or F_j should have been a good estimator of the remainder score. In fact, the associated correlations were on the order of .55 across blocks.

Two questions remain to be answered. First, can we accurately classify examinees into mastery or non-mastery states based on scores (i.e., S_j and F_j) calculated from the smaller item set? Second, was there any actual *time* savings associated with the item savings? The data relevant to the first question are reported in the next section.

Classification Analysis. Regression equations for predicting total score ($\hat{X}_j$) from both S_j and F_j were computed (Equation 2). The predicted scores ($\hat{X}_j$) were then compared to the student's observed score (X_j), and the number of correct and incorrect classifications were calculated. For both blocks the course established criterion of 70 percent was used to define the cutoff. However, using the total score as the measure of mastery or non-mastery was subject to the same criticism raised in Study I, namely that the total score is an imperfect measure of the (latent) trait of interest – mastery. The Block II and Block IV regression equations and classification analyses are presented in Table 10. As can be seen, the prediction of total score pass-fail from either S_j or F_j in Block II was almost perfect. That is, the predicted score ($\hat{X}_j$) misclassified only 1.6 percent of the sample.

In Block IV, F_j classified examinees somewhat more accurately than S_j (i.e., 97.2 percent vs. 94.4 percent). However, the errors in classification based on S_j were conservative since they classified students as failing the block test when they had actually passed.

Time Analysis. Data were collected on how long each student took to complete the flexilevel portion of the test as well as the amount of time taken to complete the remainder of the test. These times were collected for each scale in the block tests.

Table 11 presents the mean times for Blocks II and IV. The time analysis was somewhat disappointing, since the flexilevel test reduced testing time by only 15 percent and 12 percent, respectively. The procedure of starting each student at the median item of each scale required a minimum of 27 items before the flexilevel test was completed. Moreover, as pointed out earlier, those items which were not taken in the flexilevel portion tend to be the easier items and, thus answered relatively faster.

Table 10. **Regression Equations and Classification Analysis**

	Block II	Block IV
Regression Equations		
	$\hat{X}_j = .08 + .94\ S_j$	$\hat{X}_j = .03 + 1.0\ S_j$
	$\hat{X}_j = .49 + .65\ F_j$	$\hat{X}_j = .48 + .72\ F_j$

Hit-Miss Analysis Using S_j

		Block II		Block IV	
		Predicted ($\hat{X}_j$)		Predicted ($\hat{X}_j$)	
		Pass	Fail	Pass	Fail
Total Score (X_j)	Pass	52	1	57	4
	Fail	0	8	0	11
	% Correct	98.4		94.4	

Hit-Miss Analysis Using F_j

		Block II		Block IV	
		Predicted ($\hat{X}_j$)		Predicted ($\hat{X}_j$)	
		Pass	Fail	Pass	Fail
Total Score (X_j)	Pass	52	1	60	1
	Fail	0	8	1	10
	% Correct	98.4		97.2	

Table 11. **Time (in Minutes) to Complete Scales**

		Block II		Block IV
Scale	Flexilevel	Remainder	Flexilevel	Remainder
1	7.56	3.13	9.14	1.62
2	5.27	0.58	16.25	1.62
3	15.25	2.42	4.05	0.84
4	12.10	1.03	16.48	2.40
5	12.51	1.98	6.83	.93
		N = 55[a]		N = 65[a]
Total Time on Test	= 1.03 hrs			1.00 hr
Flexilevel Time	= .88 hr			.88 hr
Proportion Time Saved	= .15 hr			.12 hr

[a] Sample sizes reduced due to occasional computer failure during testing.

Conclusions

The results of the analyses suggest several conclusions about the efficacy of flexilevel testing in an ongoing training environment. First, the proportion correct during the flexilevel test (S_j) is as effective in predicting total score as the ostensibly more sensitive flexilevel score (F_j). This fact was reflected in the correlation between S_j and total score as well as the accuracy of mastery or non-mastery classification. In addition, S_j has the advantage of being in the metric that is most familiar to both students and instructors.

It was also concluded that the modest time savings (12 to 15 percent) was due to the parameters used to implement flexilevel testing. That is, entering at the median item requires the administration of at least 27 items before exit from the test. In addition, items not taken during the flexilevel test tended to be easier, as evidenced by the remainder score, which would tend to decrease the time a student needed to complete these items. However, it should be pointed out that even a 15 percent time saving applied to the large number of students in AIS courses will, in the long run, reflect an economically significant time savings.

Finally, the selection of the parameters for this study led us to speculate about potentially realizable savings due to alternate flexilevel strategies. The following study was designed to investigate that problem.

IV. STUDY III

Objective

The results of Study II were obviously contingent on the parameters chosen to implement the study. For example, examinees always began on the median item of a scale and took all scales. An alternative was to use the flexilevel algorithm at the scale level as well as the item level; i.e., if a scale is passed, the next hardest scale is taken, or, if failed, the next easiest is taken, and so on. Study I has shown that the simulation of the flexilevel algorithm on paper-and-pencil test protocols closely approximates results obtained during testing via a computer terminal. Therefore, it was decided to simulate, using Study II test protocols, the effects of adaptive movement across scales on the various dependent measures. In addition to implementing the flexilevel algorithm across scales, the simulation considered two other variables. First, the depth or item entry level within a scale was varied similar to Study I. Second, this depth notion was extended to the scale level by varying the starting scale between the hardest and easiest.

Because of the overlap in item difficulties between the original scales the items were reordered into scales based entirely on the difficulty indicies obtained in the calibration sample. The scales were formed by ranking the items according to difficulty and then forming scales with non-overlapping item difficulties. The position of a scale in the hierarchy was determined by the average difficulty of the scale. Table 12 contains the new scales for the Block II and Block IV tests.

Method

The 133 test protocols obtained during Study II were used as the data in this study. The simulation consisted of varying the levels of three parameters and measuring the effects on the various dependent measures. The three parameters manipulated were: (a) scale pass criterion (SPC), (b) scale start (SS), and (c) scale entry level (EL). These are defined as follows.

EL was used the same way as in Study I. It defined the item number within each scale where the flexilevel algorithm was started. EL was varied between 1 and 5. If EL = 1 the hardest item was given first, and if EL = 5 the 5th hardest item was given first. EL also defined the minimum number of items that had to be taken before testing within a particular scale was completed. For example, with EL = 1 at least one item had to be taken; if it was passed, testing was complete for that scale; if failed, at least one more was taken (the next easiest), and so on.

Table 12. **Items Comprising Scales and Difficulties**

Scale 1		Scale 2		Scale 3		Scale 4		Scale 5	
Item	Diff	Item	Diff	Item	Diff	Item	Diff	Item	Diff
				Block II					
15	.98	29	.94	5	.90	18	.87	35	.77
11	.97	21	.94	37	.90	8	.86	2	.75
24	.97	12	.94	3	.90	19	.85	23	.74
14	.96	31	.94	17	.89	32	.85	13	.72
9	.96	16	.93	26	.89	38	.84	28	.70
10	.96	36	.93	39	.88	22	.84	4	.70
6	.95	7	.92	25	.88	27	.81	33	.63
34	.95	20	.92			40	.81	30	.51
1	.95								
X Diff	.96		.93		.89		.84		.69
				Block IV					
15	1.00	1	.96	5	.88	27	.83	36	.69
16	1.00	8	.96	11	.88	20	.82	3	.67
18	1.00	21	.96	37	.88	22	.82	35	.62
29	1.00	38	.96	39	.88	32	.82	7	.61
26	.99	4	.95	34	.87	12	.81	6	.58
24	.98	25	.94	19	.86	28	.81	9	.58
31	.98	2	.92	14	.85	30	.72	40	.57
23	.97	10	.90	13	.84	17	.70		
						33	.70		
X Diff	.99		.94		.87		.78		.62

SS defined the scale within which testing was started, and, thus, took the values 1 through 5. If SS = 5 (the hardest scale) were passed or if SS = 1 (the easiest scale) were failed, only one scale need be taken, i.e., testing was complete.

When flexileveling at the item level, the 1-0 item score was used to define the next item to be given; i.e., a 1 implied a harder item and a 0 an easier one. In a real sense, this was the criterion for movement between items. In a similar vein, a criterion for movement between scales was needed. This was complicated by variable entry (EL), since EL = 1 implied possible scale scores of 1.0, .50, .33 and so on, whereas other values of EL implied other ranges of scale scores. Therefore, SPC was operationalized in the not wholly satisfactory sense of how many items were missed. Thus, SPC was varied between 0 and 3 where a particular value defines the maximum number of items that could be missed in order to pass the scale.

The assumption of item independence, which was important in Study I, was also relevant in this study. Namely, that a subject taking a particular item in a different order would give the same response as he gave in the original order. To the extent that this assumption is true, the results presented as follows reflect potentially obtainable outcomes from a variety of flexilevel strategies.

Results and Discussion

Simulations. The computer simulation was used to generate the values of various dependent variables for all possible combinations of the three parameters for both Block II and Block IV. The dependent variables were: (a) percent items saved, (b) percent classified correctly by S_j, (c) percent classified correctly by F_j, and (d) correlations with total score for S_j and F_j.

Table 13 presents the results of the simulation runs for Block II. Similar to Study I, EL strongly affects the dependent measures. Since EL implied the minimum number of items a student must take, the percent of items saved (% saved) varied inversely with this parameter (i.e., maximum items saved with minimum EL). Also, as EL increased, the predictiveness of S and F increased. This also was expected, since as EL increased, the item composite upon which S and F was based increased in size and thus reliability. Finally, as predictiveness increases, the percent of examinees correctly classified would be expected to increase, as it in fact does.

Table 13. **Simulation Results for Block II**

	Parameter	% Saved	Class (S)	Class (F)	Correlations	
					$R_{S,T}$	$R_{F,T}$
	0	67	.933	.932	.829	.840
	1	67	.942	.945	.833	.854
SPC[a]	2	68	.946	.948	.834	.851
	3	69	.919	.942	.830	.845
	1	63	.933	.936	.851	.872
	2	61	.949	.948	.877	.893
SS[a]	3	66	.948	.953	.859	.869
	4	71	.937	.952	.819	.829
	5	80	.908	.921	.753	.774
	1	88	.884	.883	.674	.691
	2	77	.925	.934	.817	.833
EL[a]	3	66	.954	.966	.861	.877
	4	58	.961	.966	.896	.911
	5	50	.949	.961	.911	.925

[a] Averaged over the values of the other two parameters.

The striking aspect of Table 13 was the very large savings in items obtainable with various flexilevel strategies; this was particularly dramatic for EL. At EL = 1 only 12 percent of the items were required to correctly classify nearly 90 percent of the testees. At EL = 2, only 23 percent of the original items were required to classify over 90 percent. This is in contrast to the Study II strategy, which saved 30 percent in Block II and 25 percent in Block IV, while correctly classifying 98 percent and 96 percent, respectively. It was apparent that for only a modest decrease in correct classification, an enormous increase in test items saved could be realized. If the relationship between items saved and time saved found in Study I were extrapolated to the present results, a 36 percent savings in test time could be realized at EL = 2.

The relationship of the other parameters to the dependent measures was less clear. SS would be expected to introduce a bow-shaped effect on the dependent variables, since, similar to EL, SS implies the minimum number of scales which must be taken to complete testing: SS = 3 implies at least three scales, SS = 2 or 4 implies at least 2 and SS = 1 or 5 implies at least one. This effect can be seen to some extent in the classification functions and validities – increase to SS = 2 or 3, and then decrease. Turning to SPC, there was little to choose from in terms of an optimal value. The results for SPC were perhaps idiosyncratic to the generally easy nature of the test items; i.e., varying SPC had minimal implications for all but the least prepared student.

Table 14 presents the simulation results for the Block IV test. Again, EL had the strongest effect on each dependent variable. Indeed the pattern for Block IV was much the same as the pattern reported for Block II. Looking across these blocks' results suggested that generally optimum values for the parameters were SPC = 2, SS = 3, and EL = 3.

Table 14. **Simulation Results for Block IV**

Parameter		% Saved	Class (S)	Class (F)	Correlations	
					$R_{S,T}$	$R_{F,T}$
	0	66	.895	.911	.809	.82
	1	66	.888	.919	.814	.843
SPC[a]	2	69	.886	.915	.818	.847
	3	69	.884	.900	.809	.83
	1	63	.887	.908	.823	.858
	2	60	.906	.926	.862	.883
SS[a]	3	63	.906	.926	.846	.861
	4	69	.894	.917	.812	.829
	5	79	.848	.878	.721	.749
	1	88	.853	.862	.639	.656
	2	77	.898	.910	.820	.842
EL[a]	3	65	.895	.927	.856	.882
	4	56	.895	.925	.868	.897
	5	49	.899	.931	.879	.904

[a] Averaged over the values of the other two parameters.

Table 15 presents the values of the dependent variables for the Block II and Block IV simulations using the parameter values indicated previously. These results indicated that using approximately 48 percent of the items, classified perfectly in Block II, and about 93 percent in Block IV. The correlations of both S and F with the total score were also quite high. This suggested that total score could be predicted very accurately from either score (a fact brought out by classification data).

Table 15. **Simulation Results: SPC = 2, SS = 3, EL = 3**

	% Saved	Class (S)	Class (F)	Correlations	
				$R_{S,T}$	$R_{F,T}$
Block II	54	1.00	1.00	.94	.95
Block IV	51	.93	.94	.91	.93

Since the simulation results for the two block tests were so similar, a series of regression analyses were run in order to test the generalizability of the results across blocks. Basically, the results presented to this point addressed the question of what kinds of item savings and classification accuracy could be expected by various flexilevel strategies. The overriding question, however, was the extent to which these results generalize from block to block. If the simulation results, or for that matter the empirical results from Study II, are block specific (i.e., content or item characteristic specific), then they are of little value in forecasting what would happen in a new block. If the results show generalizability across the two blocks of PME, there is evidence that implementing a particular flexilevel testing strategy in a new block, or even a new course with similar item characteristics, would yield similar outcomes.

Regression Analyses. The predictability of the dependent variables generated in the simulation studies was assessed in a number of regression analyses. The predictor variables were the original parameters SS, SPC and EL, plus certain nonlinear predictors. These latter predictors were derived from the original three parameters and were EL², EL x SPC, 1n(EL), 1n (EL) x SS, EL x SS, SS², SS³ and |SS – 3|.

The inclusion of these derived variables produced a total of nine predictors. Regression runs were done separately by block for: (a) percent items saved, (b) percent correctly classified by S (Class (S)), (c) percent correctly classified by F (Class (F)), (d) the constant term for predicting the total score from the S score (a_S), (e) the b weight for predicting total score from the S score (b_S), (f) the constant for predicting the total score from the F score (a_F), and (g) the b weight for predicting the total score from the F score (b_F). In addition, regression analyses were run after combining Block II and Block IV data.

The Block II, Block IV, and total regression analyses for the percent items saved criterion are presented in Table 16. As can be seen percent items saved was highly predictable in all three analyses, with EL having the greatest weight. This was consistent with the results presented in Tables 13 and 14, and was highly consistent across blocks, as well as when the block data were pooled. This was reflected in the multiple correlations for each analysis as well as the consistency of the beta weights across the three analyses.

Table 16. **Regression Analyses for % Items Saved**

	Block II	Block IV	Total
Variable	Beta weights	Beta weights	Beta weights
EL	–1.35	–1.35	–1.34
SS³	.24	.24	.24
EL x SS	.35	.22	.28
|SS–3|	.10	.11	.11
EL²	.21	.28	.25
EL x SPC	.07	.11	.09
Multiple R	.96	.93	.94
R Square	.92	.87	.89

Table 17 presents the three regression analyses using Class(S) as the criterion. These analyses were not as consistent as those reported in Table 16. Again, EL or some transformation of EL was the most important variable in predicting the classification power of S_j. Class(S) was not as predictable as the percent items saved as evidenced by the relatively low multiple R's in comparison to those reported in Table 16. Furthermore, a different set of predictors was defined for each analysis, however, the relative ranking of the common predictors (viz, EL, SS², and EL²) was the same across all three analyses.

Table 17. **Regression Analyses for Predicting Class(S)**

	Block II	Block IV	Total
Variable	Beta weights	Beta weights	Beta weights
EL	1.34	1.75	1.29
SS²	–.63	–.36	–.26
EL²	–1.33	–.67	–.94
|SS–3|		–.28	–.21
1n(EL)		–.83	
EL x SS	.68		
SPC	–.10		
Multiple R	.81	.68	.60
R Square	.66	.46	.36

Table 18 presents the analyses using class (F) as the criterion variable. Again, EL was the variable which had the largest beta weight followed by EL^2. In this analysis the four common variables retained their relative importance across the blocks and the total sample analyses. The multiple correlations were somewhat higher than those obtained in the Class (S) analysis especially for the total sample.

Table 18. **Regression Analyses for Predicting Class (F)**

	Block II	Block IV	Total
Variable	Beta weights	Beta weights	Beta weights
EL	1.42	1.45	1.39
EL^2	−1.48	−.99	−1.16
|SS−3|	−.09	−.24	−.16
EL x SPC	.17		
SS^3	−.49	−.26	−.36
EL x SS	.69		.33
Multiple R	.87	.70	.71
R Square	.75	.49	.51

Table 19 contains analyses for a_S. As evidenced by the multiple R's, a_S is almost perfectly predictable. As was the case in all other analyses, EL was the most predictive.

Table 19. **Regression Analyses Predicting a_s**

	Block II	Block IV	Total
Variable	Beta weights	Beta weights	Beta weights
EL	-1.46	−1.59	−1.83
SS^2	.47	.33	.35
EL x SPC	.16	.30	.24
1n(EL) x SS	.25		−.24
EL^2	.41	.12	.25
|SS−3|	−.06	−.06	−.05
SPC	.11	.08	.09
EL x SS		.44	.60
1n(EL)		.20	.43
Multiple R	.99	.99	.98
R Square	.99	.98	.96

Table 20 contains the regression analyses using b_S as the criterion. Again the most important predictor in all three analyses was EL. The prediction for all three analyses was essentially perfect with the lowest multiple correlation coefficient being .98. This suggests that the b weight for the S score in predicting the total score was highly predictable from the three parameters studied in the simulations. The results of the a_S analyses also suggest that the constant term in the regression equation is highly predictable (see Table 19).

Table 20. **Regression Analyses Predicting b_s**

Variable	Block II Beta weights	Block IV Beta weights	Total Beta weights
EL	2.18	1.41	1.90
SS^3	–.40	–.38	–.37
EL x SPC	–.14	–.25	–.20
ln(EL) x SS	.38	–.28	.20
EL^2	–.52	–.24	–.36
|SS–3|		.06	.03
ln(EL)	–.6		–.45
EL x SS	–.59		–.45
Multiple R	.99	.985	.98
R Square	.99	.97	.97

Tables 21 and 22 present the results for a_F and b_F, respectively. While a_F was highly predictable within blocks, it was not so predictable in the pooled sample. This suggested that the regressions were not homogeneous and, therefore, the parameter was specific to test content or item characteristics. In contrast, the b_F analysis (Table 22) shows remarkable consistency, both across blocks and when pooled.

Table 21. **Regression Analyses Predicting a_F**

Variable	Block II Beta weights	Block IV Beta weights	Total Beta weights
SS^3	–.99	–.82	–.40
EL	.57	.52	.13
|SS–3|	.22	.31	.12
ln(EL) x SS		.13	
EL^2	–.56		
ln(EL)	–.20		
Multiple R	.97	.96	.40
R Square	.94	.92	.16

Table 22. **Regression Analyses for b_F**

Variable	Block II Beta weights	Block IV Beta weights	Total Beta weights
EL	1.94	1.63	1.84
EL^2	–.59	–.37	–.49
EL x SPC	–.08	–.18	–.09
SPC		.08	
ln(EL)	–.58	–.33	–.54
ln(EL) x SS	.30	–.16	.17
EL x SS	–.41		–.31
SS^3	–.62	–.59	–.59
|SS–3|		.08	.04
Multiple R	.993	.978	.983
R Square	.987	.956	.966

Cross-Validation. To assess the generality of the regression equations they were cross-validated by using the Block II and total weights to predict in Block IV. Similarily, the weights obtained in Block IV and total analyses were cross-validated against Block II data. The results of the cross-validation analyses are presented in Table 23. As can be seen, very little shrinkage occurred for the majority of the variables under study. The greatest shrinkage occurred for a_F when either block was validated against the other.

Table 23. **Multiple Correlations Obtained in Cross-Validation Study**

	Block II			Block IV		
Criterion	Original Multiple R	CV4 Multiple R	CVT Multiple R	Original	CV2	CVT
Save	.96	.96	.96	.93	.93	.93
Class (S)	.81	.67	.77	.68	.54	.63
Class (F)	.88	.79	.84	.70	.60	.68
a_S	1.00	.98	.99	1.00	.98	.99
b_S	1.00	.98	.99	.99	.98	.98
a_F	.97	.62	.85	.96	.63	.88
b_F	.99	.97	.99	.98	.97	.98

Conclusions and Recommendations

Study III has shown that large savings in items, and, potentially, test time, can be realized through the implementation of alternate flexilevel strategies. The conservative strategy adopted in Study II resulted in only modest item and time savings. However, even these modest savings can result in significant dollar savings when amortized over the thousands of technical training students for just one year. Study III has shown that significantly greater savings can be realized with more efficient procedures, in the form of optimal values for SPC, SS, and EL.

More important, the cross-validation results of Table 23 suggest that for testing situations similar to those studied here, a course designer can plan the implementation of flexilevel testing with considerable accuracy. After making a determination of the appropriate strategy; i.e., selecting levels of SPC, SS, and EL, the planner can estimate the amount of item savings which will occur and can make initial estimates of total score. The latter can be accomplished by substituting the selected parameters into any of the equations in Table 19 to obtain a; then any of the equations in Table 20 to obtain b. These weights, then, are the regression parameters to convert S_j, the percent of items correct, to an estimated total percent correct score.

Table 23 offers some evidence those parameters discussed above, percent items saved, a_S and b_S, are estimative independent of block content, since the equations are virtually interchangeable between the blocks studied in this research and highly predictive as well. With the exception of b_F, the other dependent measures in Table 23 are predictable in varying degrees, often with significant shrinkage. As mentioned earlier, shrinkage is an indication that the outcome measures are more a function of the test content or item characteristics, than of the testing strategy. In these instances it is important to develop estimates which are *specific to the particular testing situation. In any event, caution would dictate that any newly implemented* flexilevel testing program be validated to determine its efficiency.

The overall conclusion from the three studies would seem to be that flexilevel testing, with variable entry, offers an easily implemented testing procedure with potential for significant dollar savings at minimal risk (in the sense of misclassification). Studies I and III, the simulation studies, show the potential power of implementing alternate strategies and the great regularity of the data obtained.

Study I indicated the viability of simulating flexilevel testing on paper-and-pencil protocols to determine optimal entry levels as well as potential item savings. This type of simulation can be accomplished prior to actual implementation, or the results from Study III can be used directly to guide the selection of an optimal flexilevel strategy.

In any event, it would seem appropriate to implement further flexilevel testing in technical training where the availability of computer terminals permits. Since, for example, in the Advanced Instructional System, students spend 30 to 40 percent of their time in testing activities, it can be seen that significant training time reductions are potentially obtainable.

REFERENCES

Bryson, R. Shortening tests: Effects of method used, length and internal consistency on correlation with total score. *Proceedings of the 80th Annual Convention of the American Psychological Association*, 1972, 7-8.

Cleary, T.A., Linn, R.L., & Rock, D.A. An exploratory study of programmed tests. *Educational and Psychological Measurement*, 1968, **28**, 345–360. (a)

Cleary, T.A., Linn, R.L., & Rock, D.A. Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement*, 1968, **5**, 183-187. (b)

Hansen, D.N., Harris, D.A., & Ross, S. *Flexilevel adaptive testing paradigm: Validation in technical training.* AFHRL-TR-77-35(I), AD-A042 977. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, July 1977. (a)

Hansen, D.N., Harris, D.A., & Ross, S. *Flexilevel adaptive testing paradigm: Hierarchical concept structures.* AFHRL-TR-77-35(II), AD-A042 966. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, July 1977. (b)

Hansen, D.N., Johnson, B.F., Fagan, R.L., Tam, P., & Dick, W. *Computer-based adaptive testing models for the Air Force technical training environment Phase I: Development of a computerized measurement system for Air Force technical training.* AFHRL-TR-74-48, AD-785 142. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, July 1974.

Linn, R.L., Rock, D.A., & Cleary, T.A. *Sequential testing for dichotomous decisions.* College Entrance Examination Board Research and Development Report. RDR 60-80, No. 3, 1970 (ETS, Rb-70, 31).

Lord, F.M. The self-scoring flexilevel test. *Journal of Educational Measurement*, 1971, **8**, 147-151. (a)

Lord, F.M. A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 1971, **31**, 808-813. (b)

Patterson, J.J. *An evaluation of the sequential method of psychological testing.* Unpublished doctoral dissertation, Michigan State University, 1962.

☆U.S. GOVERNMENT PRINTING OFFICE: 1978–771-122/85